Lecture 11 presents the protein structure prediction method of comparative modeling. The idea is that for a target sequence which lacks a 3D structure, a comparative protein model can be built at atomic resolution using one or more known structures. The known structures that are used to build the target (the unknown) are reference as templates. For high SeqID (roughly > 25-30%), the templates can be found via a Blast search of the target sequence against sequences from known structures (using PDB). For low SeqID, fold recognition is applied to find the structural templates (Lecture 10). If all fails, then ab initio methods are applied to predict a fold (e.g., ITASSER or Rosetta using fragments to construct a protein model).

An important element of comparative protein modeling is sequence-sequence alignment of the target vs. the known structure. In the case of high SeqID, the alignment can be achieved with good precision. For fold recognition, it is quite often that protein folds are recognized to good accuracy but the sequence alignment is poor. It is easier to detect a fold at the global level, yet miss the details at the residue level and its correspondence in an alignment with known structures. A misaligned sequence to a structure ultimately leads to an incorrect protein model being predicted. However, in principle, improvement in the model can be obtained by “structure refinement;" -- however, current computational methods that are relieable are still lacking (a hot topic of research).

Discussion topics.

Topic 1
Find a protein target sequence that lacks a crystallographic or NMR structure, yet the sequence is a homolog to one or more known structures. The target should be greater than 200 residues in length and a member of a family (or superfamily) where some of the members are known at the structural level. There is a wealth of examples (e.g., members of the smallpox family, etc). Set the sequence identity to a threshold of 30% or greater. Provide the following:

Describe your selected protein.
Use SwissModel to predict a 3D model of your selected protein. List the template and SeqID. Is there complete coverage of your target sequence with the structural template?
Apply the ESyPred3D Web Server (or, if not working correctly, select a different method from https://en.wikipedia.org/wiki/List_of_protein_structure_prediction_software) and I-TASSER to predict 3D models of your selected protein, using the exact template found from SwissModel. (Check each server for input of PDB template -- in I-TASSER, use the Option 1 to select template.) From I-TASSER calculations, use the rank-1 hit as your predicted model. Are the sequence-sequence alignments used to build the models the same among the three servers? If not, explain.
Use structure-structure alignment methods to align the predicted 3D models from SwissModel, ESyPred3D and I-TASSER. Report values of RMSD and SeqID. Are the values of RMSD = 0 and the SeqID = 100%? Explain any deviation.
Topic 2
Let us imagine that a scientific journal asks you to referee a manuscript submitted for publication in which the paper describes a comparative protein model and its application in

early-stage drug discovery. What reported details of the protein model would you require to accept the manuscript for publication? Act as a referee and list your concerns about the accuracy and reliability of the model and its application to structure-based drug development.

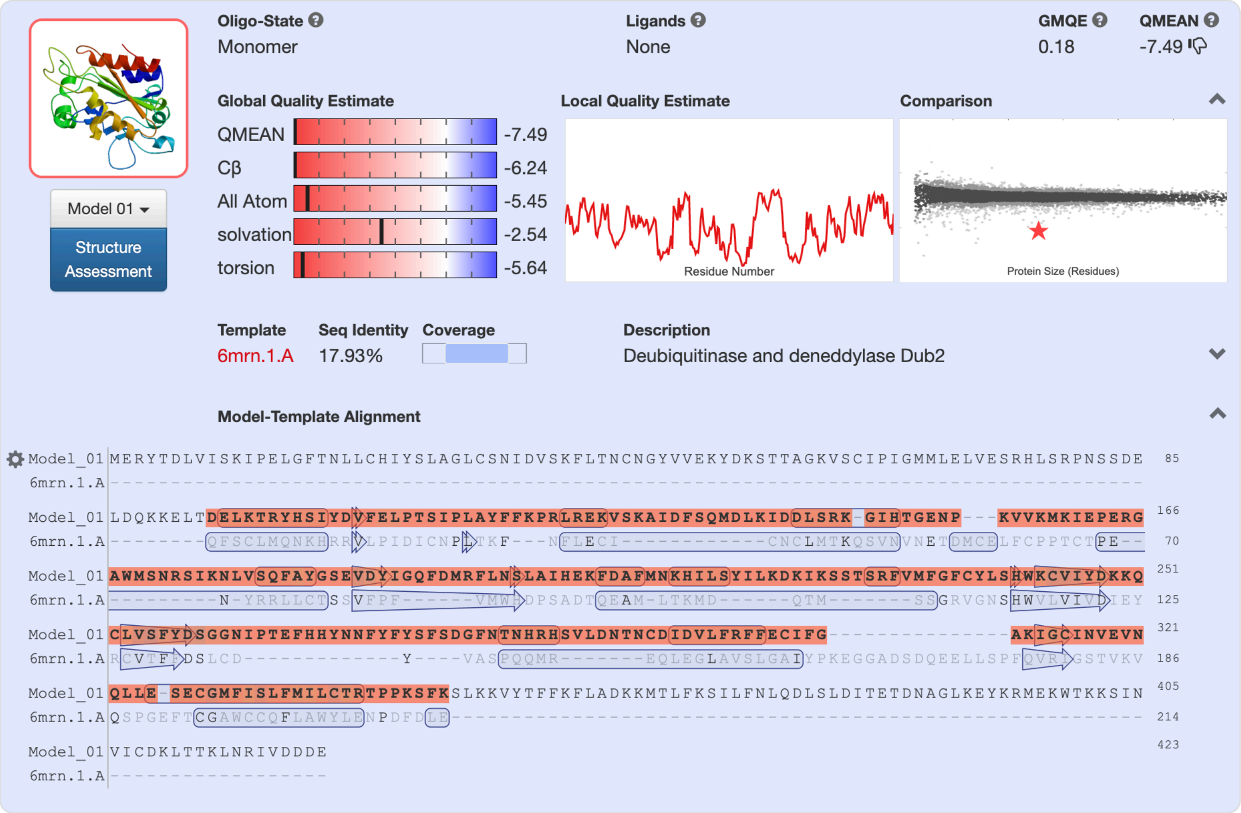
https://www.uniprot.org/uniprot/Q0NG77

Organism: Variola Virus (smallpox)

Protein: Viral Core cysteine proteinase

```
>tr|Q0NG77|Q0NG77_VAR46 Viral core cysteine proteinase OS=Variola virus
(isolate Human/Japan/Yamada MS-2(A)/1946) OX=587202 GN=VARV_JAP46_yam_065
PE=4 SV=1
MERYTDLVISKIPELGFTNLLCHIYSLAGLCSNIDVSKFLTNCNGYVVEKYDKSTTAGKV
SCIPIGMMLELVESRHLSRPNSSDELDQKKELTDELKTRYHSIYDVFELPTSIPLAYFFK
PRLREKVSKAIDFSQMDLKIDDLSRKGIHTGENPKVVKMKIEPERGAWMSNRSIKNLVSQ
FAYGSEVDYIGQFDMRFLNSLAIHEKFDAFMNKHILSYILKDKIKSSTSRFVMFGFCYLS
HWKCVIYDKKQCLVSFYDSGGNIPTEFHHYNNFYFYSFSDGFNTNHRHSVLDNTNCDIDV
LFRFFECIFGAKIGCINVEVNQLLESECGMFISLFMILCTRTPPKSFKSLKKVYTFFKFL
ADKKMTLFKSILFNLQDLSLDITETDNAGLKEYKRMEKWTKKSINVICDKLTTKLNRIVD
DDE
```

The protein I chose to look at was the viral core cysteine proteinase from the Variola virus, or small pox. This particular protein was isolated from a human host from Japan. It has a length of 423 amino acids and shared identities with core protease I7. The core protease I& is a protein that is in charge of processing viral core and membrane proteins to undergo proteolysis, which is involved in transforming immature virions to mature virions. The template ID from Swiss-Model is 6mrn.1.A with a seq ID of 17.93%.



Oligo-State	Ligands	GMQE	QMEAN
Monomer	None	0.18	-7.49

Model 01

Structure Assessment

Global Quality Estimate

QMEAN	-7.49
Cβ	-6.24
All Atom	-5.45
solvation	-2.54
torsion	-5.64

Local Quality Estimate

Comparison

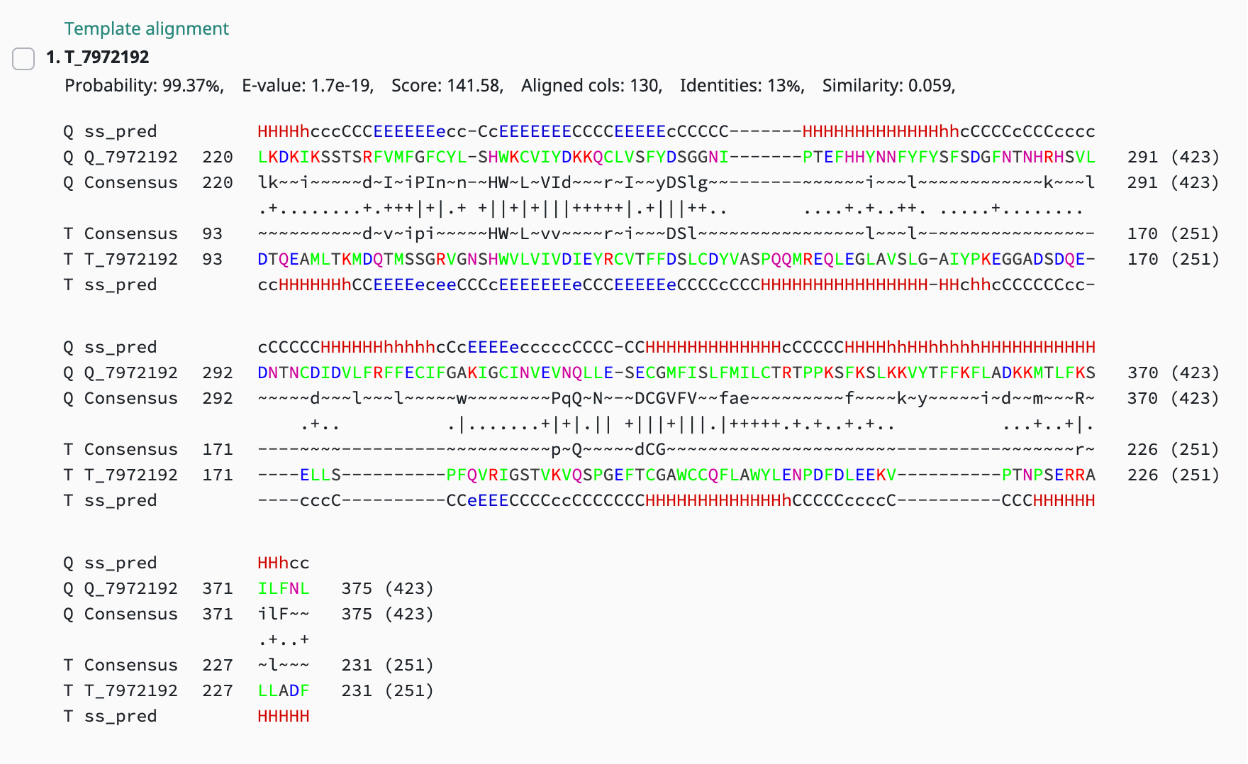
Template	Seq Identity	Coverage	Description
6mrn.1.A	17.93%		Deubiquitinase and deneddylase Dub2

Model-Template Alignment

```
Model_01 MERYTDLVISKIPELGFTNLLCHIYSLAGLCSNIDVSKFLTNCNGYVVEKYDKSTTAGKVSCIPIGMMLELVESRHLSRPNSSDE  85
6mrn.1.A -------------------------------------------------------------------------------------

Model_01 LDQKKELTDELKTRYHSIYDVFELPTSIPLAYFFKPRLREKVSKAIDFSQMDLKIDDLSRK-GIHTGENP---KVVKMKIEPERG 166
6mrn.1.A --------QFSCLMQNKHRRVLPIDICNPLTKF---NFLECI------------CNCLMTKQSVNVNETDMCELFCPPTCTPE--  70

Model_01 AWMSNRSIKNLVSQFAYGSEVDYIGQFDMRFLNSLAIHEKFDAFMNKHILSYILKDKIKSSTSRFVMFGFCYLSHWKCVIYDKKQ 251
6mrn.1.A ---------N-YRRLLCTSSVFPF------VMWHDPSADTQEAM-LTKMD------QTM--------SSGRVGNSHWVLVIVDIEY 125

Model_01 CLVSFYDSGGNIPTEFHHYNNFYFYSFSDGFNTNHRHSVLDNTNCDIDVLFRFFECIFG---------------AKIGCINVEVN 321
6mrn.1.A RCVTFFDSLCD-------------Y----VASPQQMR-------EQLEGLAVSLGAIYPKEGGADSDQEELLSPFQVRIGSTVKV 186

Model_01 QLLE-SECGMFISLFMILCTRTPPKSFKSLKKVYTFFKFLADKKMTLFKSILFNLQDLSLDITETDNAGLKEYKRMEKWTKKSIN 405
6mrn.1.A QSPGEFTCGAWCCQFLAWYLENPDFDLE--------------------------------------------------------- 214

Model_01 VICDKLTTKLNRIVDDDE                                                                    423
6mrn.1.A ------------------
```

Unfortunately, Esypred was not allowing me to submit any jobs, so I decided to use HHpred to align my sequence of interest to that of 6mrn_A found as the template from my SWISS-Model run.

HHPred



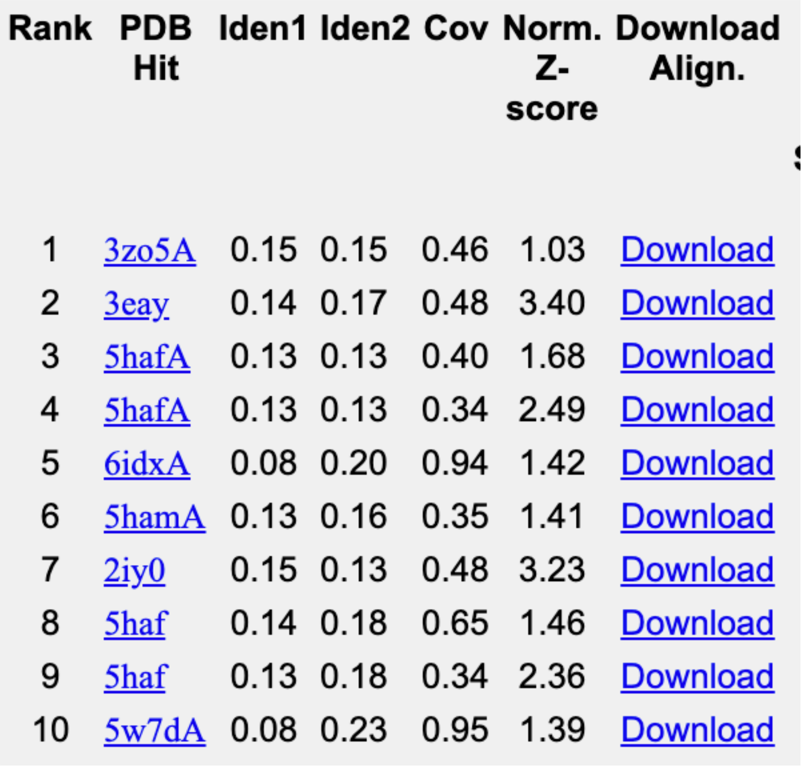
□ **1. T_7972192**

Probability: 99.37%,   E-value: 1.7e-19,   Score: 141.58,   Aligned cols: 130,   Identities: 13%,   Similarity: 0.059,

```
Q ss_pred              HHHHhcccCCCEEEEEEecc-CcEEEEEEECCCCEEEEEcCCCCC-------HHHHHHHHHHHHHhhcCCCCcCCCcccc
Q Q_7972192    220     LKDKIKSSTSRFVMFGFCYL-SHWKCVIYDKKQCLVSFYDSGGNI-------PTEFHHYNNFYFYSFSDGFNTNHRHSVL    291 (423)
Q Consensus    220     lk~~i~~~~~d~I~iPIn~n-~HW~L~VId~~~r~I~~yDSlg~~-------~~~~~~i~~~l~~~~~~~~~~~~k~~~l    291 (423)
                       .+........+.+++|+|.+ +||+|+||||+++++|.+|||++..       ....+.+..++. .....+........
T Consensus    93      ~~~~~~~~~~d~v~ipi~~~~~HW~L~vv~~~~r~i~~~DSl~~~~~~~~~~~~~~~~~~l~~~l~-~~~~~~~~~~~~~~-    170 (251)
T T_7972192    93      DTQEAMLTKMDQTMSSGRVGNSHWVLVIVDIEYRCVTFFDSLCDYVASPQQMREQLEGLAVSLG-AIYPKEGGADSDQE-    170 (251)
T ss_pred              ccHHHHHHhCCEEEEeceeCCCcEEEEEEEeCCCEEEEEeCCCCcCCCHHHHHHHHHHHHHHHH-HHchhcCCCCCCcc-


Q ss_pred              cCCCCCHHHHHHhhhhhcCcEEEEecccccCCCC-CCHHHHHHHHHHHHHcCCCCCHHHHhhHHhhhhhHHHHHHHHHHH
Q Q_7972192    292     DNTNCDIDVLFRFFECIFGAKIGCINVEVNQLLE-SECGMFISLFMILCTRTPPKSFKSLKKVYTFFKFLADKKMTLFKS    370 (423)
Q Consensus    292     ~~~~~d~~~l~~~l~~~~~w~~~~~~~~~~PqQ~N~-~DCGVFV~~fae~~~~~~~~~~f~~~~k~y~~~~~i~d~~m~~~R~    370 (423)
                           .+..          .|.......+|+|.|| +|||+|||.|+++++.+.+..+.+..          ...+..+|.
T Consensus    171     ----~~~~----------~~~~~~~~~~p~Q~~~~~dCG~~~~~~~~~~~~~~~~~~~~~~~~~----------~~~~~~~r~    226 (251)
T T_7972192    171     ----ELLS----------PFQVRIGSTVKVQSPGEFTCGAWCCQFLAWYLENPDFDLEEKV----------PTNPSERRA    226 (251)
T ss_pred              ----cccC----------CCeEEECCCCccCCCCCCCHHHHHHHHHHHHHhCCCCCccccC----------CCCHHHHHH


Q ss_pred              HHhcc
Q Q_7972192    371     ILFNL    375 (423)
Q Consensus    371     ilF~~    375 (423)
                       .+..+
T Consensus    227     ~l~~~    231 (251)
T T_7972192    227     LLADF    231 (251)
T ss_pred              HHHHH
```

I-TASSER

The top PDB hit for I-TASSER was 3zo5A.



Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score	Download Align.
1	3zo5A	0.15	0.15	0.46	1.03	Download
2	3eay	0.14	0.17	0.48	3.40	Download
3	5hafA	0.13	0.13	0.40	1.68	Download
4	5hafA	0.13	0.13	0.34	2.49	Download
5	6idxA	0.08	0.20	0.94	1.42	Download
6	5hamA	0.13	0.16	0.35	1.41	Download
7	2iy0	0.15	0.13	0.48	3.23	Download
8	5haf	0.14	0.18	0.65	1.46	Download
9	5haf	0.13	0.18	0.34	2.36	Download
10	5w7dA	0.08	0.23	0.95	1.39	Download

Running the ITASSER result of 3zo5A and the SWISS-Model result of 6mrn through FATCAT alignment, I found that the two structures have a RMSD of 2.88 and a Seq ID of 7.87%.